Lectures 2: Review of Pipelines and Caches

Prof. David A. Patterson
Computer Science 252
Fall 1996
Review, #1

• Designing to Last through Trends

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic</td>
<td>2x in 3 years</td>
</tr>
<tr>
<td></td>
<td>2x in 3 years</td>
</tr>
<tr>
<td>DRAM</td>
<td>4x in 3 years</td>
</tr>
<tr>
<td></td>
<td>1.4x in 10 years</td>
</tr>
<tr>
<td>Disk</td>
<td>4x in 3 years</td>
</tr>
<tr>
<td></td>
<td>1.4x in 10 years</td>
</tr>
</tbody>
</table>

• Time to run the task
  – Execution time, response time, latency

• Tasks per day, hour, week, sec, ns, ...
  – Throughput, bandwidth

• “X is n times faster than Y” means

\[
\frac{\text{ExTime}(Y)}{\text{ExTime}(X)} = \frac{\text{Performance}(X)}{\text{Performance}(Y)}
\]
Review, #2

• Amdahl’s Law:

\[
\text{Speedup}_{\text{overall}} = \frac{\text{ExTime}_{\text{old}}}{\text{ExTime}_{\text{new}}} = \frac{1}{(1 - \text{Fraction}_{\text{enhanced}}) + \frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}}}
\]

• CPI Law:

<table>
<thead>
<tr>
<th>CPU time</th>
<th>= Seconds</th>
<th>Instructions \times Cycles \times Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>Program</td>
<td>Instruction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cycle</td>
</tr>
</tbody>
</table>

• Execution time is the REAL measure of computer performance!

• Good products created when have:
  – Good benchmarks
  – Good ways to summarize performance

• Die Cost goes roughly with die area$^4$
Review, #3: Price vs. Cost

- **Mini W/S PC**
  - Average Discount: 4.7
  - Gross Margin: 3.5
  - Direct Costs: 2.5
  - Component Costs: 1.8

- **W/ S**
  - Average Discount: 3.8
  - Gross Margin: 2.5
  - Direct Costs: 1.8
  - Component Costs: 1.5
Pipelining: Its Natural!

• Laundry Example
• Ann, Brian, Cathy, Dave each have one load of clothes to wash, dry, and fold
• Washer takes 30 minutes
• Dryer takes 40 minutes
• “Folder” takes 20 minutes
Sequential Laundry

• Sequential laundry takes 6 hours for 4 loads
• If they learned pipelining, how long would laundry take?
Pipelined Laundry
Start work ASAP

- Pipelined laundry takes 3.5 hours for 4 loads
Pipelining Lessons

- Pipelining doesn’t help latency of single task, it helps throughput of entire workload
- Pipeline rate limited by slowest pipeline stage
- Multiple tasks operating simultaneously
- Potential speedup = Number pipe stages
- Unbalanced lengths of pipe stages reduces speedup
- Time to “fill” pipeline and time to “drain” it reduces speedup
Computer Pipelines

- Execute billions of instructions, so throughout is what matters
- DLX desirable features: all instructions same length, registers located in same place in instruction format, memory operands only in loads or stores
5 Steps of DLX Datapath

Figure 3.1, Page 130
Pipelined DLX Datapath

Figure 3.4, page 137

- Data stationary control
  - local decode for each instruction phase / pipeline stage
Visualizing Pipelining

Figure 3.3, Page 133

Time (clock cycles)
Its Not That Easy for Computers

• Limits to pipelining: **Hazards** prevent next instruction from executing during its designated clock cycle
  – **Structural hazards**: HW cannot support this combination of instructions (single person to fold and put clothes away)
  – **Data hazards**: Instruction depends on result of prior instruction still in the pipeline (missing sock)
  – **Control hazards**: Pipelining of branches & other instructions that change the PC (football uniform analogy)

• Common solution is to **stall** the pipeline until the hazard is resolved, inserting one or more “**bubbles**” in the pipeline
One Memory Port/Structural Hazards

Figure 3.6, Page 142

Time (clock cycles)

Instr Order

Instr 1
Instr 2
Instr 3
Instr 4

Load
One Memory Port/Structural Hazards

Figure 3.7, Page 143

Time (clock cycles)

Instr. Order

Instr 1

Instr 2

Instr 3

Load

Instr 3 stall
Speed Up Equation for Pipelining

\[ CPI_{\text{pipelined}} = \text{Ideal CPI} + \text{Pipeline stall clock cycles per instr} \]

\[ \text{Speedup} = \frac{\text{Ideal CPI} \times \text{Pipeline depth}}{\text{Ideal CPI} + \text{Pipeline stall CPI}} \times \frac{\text{Clock Cycle}_{\text{unpipelined}}}{\text{Clock Cycle}_{\text{pipelined}}} \]

\[ \text{Speedup} = \frac{\text{Pipeline depth}}{1 + \text{Pipeline stall CPI}} \times \frac{\text{Clock Cycle}_{\text{unpipelined}}}{\text{Clock Cycle}_{\text{pipelined}}} \]
Example: Dual-port vs. Single-port

- Machine A: Dual ported memory
- Machine B: Single ported memory, but its pipelined implementation has a 1.05 times faster clock rate
- Ideal CPI = 1 for both
- Loads are 40% of instructions executed

\[
\text{SpeedUp}_A = \frac{\text{Pipeline Depth}}{1 + 0} \times \frac{\text{clock}_{\text{unpipe}}}{\text{clock}_{\text{pipe}}} \\
= \text{Pipeline Depth}
\]

\[
\text{SpeedUp}_B = \frac{\text{Pipeline Depth}}{1 + 0.4 \times 1} \times \frac{\text{clock}_{\text{unpipe}}}{\text{clock}_{\text{unpipe}}/1.05} \\
= \frac{\text{Pipeline Depth}}{1.4} \times 1.05 \\
= 0.75 \times \text{Pipeline Depth}
\]

\[
\frac{\text{SpeedUp}_A}{\text{SpeedUp}_B} = \frac{\text{Pipeline Depth}}{(0.75 \times \text{Pipeline Depth})} = 1.33
\]

- Machine A is 1.33 times faster
Data Hazard on R1

Figure 3.9, page 147

Time (clock cycles)

add r1, r2, r3

sub r4, r1, r3

and r6, r1, r7

or r8, r1, r9

xor r10, r1, r11
Three Generic Data Hazards

Instr\textsubscript{i} followed by Instr\textsubscript{j}

• Read After Write (RAW)
  Instr\textsubscript{j} tries to read operand before Instr\textsubscript{i} writes it
Three Generic Data Hazards

Instr\_i followed by Instr\_j

- **Write After Read (WAR)**
  Instr\_j tries to write operand *before* Instr\_i reads it

- Can’t happen in DLX 5 stage pipeline because:
  - All instructions take 5 stages,
  - Reads are always in stage 2, and
  - Writes are always in stage 5
Three Generic Data Hazards

Instr\textsubscript{i} followed by Instr\textsubscript{j}

- **Write After Write (WAW)**
  Instr\textsubscript{j} tries to write operand *before* Instr\textsubscript{i} writes it
  - Leaves wrong result (Instr\textsubscript{i} not Instr\textsubscript{j})

- Can’t happen in DLX 5 stage pipeline because:
  - All instructions take 5 stages, and
  - Writes are always in stage 5

- Will see WAR and WAW in later more complicated pipes
CS 252 Administrivia

• Students with too varied background?
  – In past, CS grad students took written prelim exams on undergraduate material in hardware, software, and theory
  – Prelims were dropped => some unprepared for CS 252?

• In class exam on Wednesday September 3
  – Bring up to 2 sheets of paper with notes on both sides
  – Doesn’t affect grade, only admission into class
  – 2 grades: Admitted or audit/take CS 152 1st (same time in 306)
  – Improve your experience if recapture common background

• Review: Chapters 1-3, CS 152 home page, maybe “Computer Organization and Design (COD)”
  – Chapters 1 to 8 of COD if never took prerequisite
  – If did take a class, be sure COD Chapters 2, 6, 7 are familiar
  – Copies in Bechtel Library on 2-hour reserve
Forwarding to Avoid Data Hazard

Figure 3.10, Page 149

Time (clock cycles)

Instr. Order

add r1, r2, r3
sub r4, r1, r3
and r6, r1, r7
or  r8, r1, r9
xor r10, r1, r11
HW Change for Forwarding

Figure 3.20, Page 161
Data Hazard Even with Forwarding

Figure 3.12, Page 153

 lw r1, 0(r2)
 sub r4, r1, r6
 and r6, r1, r7
 or r8, r1, r9
Data Hazard Even with Forwarding

Figure 3.13, Page 154

Instruction Order

Time (clock cycles)

lw r1, 0(r2)
sub r4, r1, r6
and r6, r1, r7
or r8, r1, r9
Software Scheduling to Avoid Load Hazards

Try producing fast code for

\[ a = b + c; \]
\[ d = e - f; \]

assuming \( a, b, c, d, e, \) and \( f \) in memory.

Slow code:

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Slow code</th>
<th>Fast code</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW</td>
<td>Rb,b</td>
<td>LW Rb,b</td>
</tr>
<tr>
<td>LW</td>
<td>Rc,c</td>
<td>LW Rc,c</td>
</tr>
<tr>
<td>ADD</td>
<td>Ra,Rb,Rc</td>
<td>LW Re,e</td>
</tr>
<tr>
<td>SW</td>
<td>a,Ra</td>
<td>ADD Ra,Rb,Rc</td>
</tr>
<tr>
<td>LW</td>
<td>Re,e</td>
<td>LW Rf,f</td>
</tr>
<tr>
<td>LW</td>
<td>Rf,f</td>
<td>SW a,Ra</td>
</tr>
<tr>
<td>SUB</td>
<td>Rd,Re,Rf</td>
<td>SUB Rd,Re,Rf</td>
</tr>
<tr>
<td>SW</td>
<td>d,Rd</td>
<td>SW d,Rd</td>
</tr>
</tbody>
</table>
Control Hazard on Branches
Three Stage Stall
Branch Stall Impact

• If CPI = 1, 30% branch, Stall 3 cycles => new CPI = 1.9!
• Two part solution:
  – Determine branch taken or not sooner, AND
  – Compute taken branch address earlier
• DLX branch tests if register = 0 or ≠ 0
• DLX Solution:
  – Move Zero test to ID/RF stage
  – Adder to calculate new PC in ID/RF stage
  – 1 clock cycle penalty for branch versus 3
Pipelined DLX Datapath

Figure 3.22, page 163


This is the correct 1 cycle latency implementation!
Four Branch Hazard Alternatives

#1: Stall until branch direction is clear

#2: Predict Branch Not Taken
- Execute successor instructions in sequence
- “Squash” instructions in pipeline if branch actually taken
- Advantage of late pipeline state update
- 47% DLX branches not taken on average
- PC+4 already calculated, so use it to get next instruction

#3: Predict Branch Taken
- 53% DLX branches taken on average
- But haven’t calculated branch target address in DLX
  » DLX still incurs 1 cycle branch penalty
  » Other machines: branch target known before outcome
Four Branch Hazard Alternatives

#4: Delayed Branch

- Define branch to take place AFTER a following instruction

\[
\text{branch instruction} \\
\text{sequential successor}_1 \\
\text{sequential successor}_2 \\
\vdots \\
\text{sequential successor}_n \\
\text{branch target if taken}
\]

- 1 slot delay allows proper decision and branch target address in 5 stage pipeline
- DLX uses this

Branch delay of length \( n \)
Delayed Branch

• Where to get instructions to fill branch delay slot?
  – Before branch instruction
  – From the target address: only valuable when branch taken
  – From fall through: only valuable when branch not taken
  – Cancelling branches allow more slots to be filled

• Compiler effectiveness for single branch delay slot:
  – Fills about 60% of branch delay slots
  – About 80% of instructions executed in branch delay slots useful in computation
  – About 50% (60% x 80%) of slots usefully filled
Evaluating Branch Alternatives

\[
\text{Pipeline speedup} = \frac{\text{Pipeline depth}}{1 + \text{Branch frequency} \times \text{Branch penalty}}
\]

<table>
<thead>
<tr>
<th>Scheduling scheme</th>
<th>Branch penalty</th>
<th>CPI</th>
<th>speedup v. unpipelined</th>
<th>speedup v. stall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stall pipeline</td>
<td>3</td>
<td>1.42</td>
<td>3.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Predict taken</td>
<td>1</td>
<td>1.14</td>
<td>4.4</td>
<td>1.26</td>
</tr>
<tr>
<td>Predict not taken</td>
<td>1</td>
<td>1.09</td>
<td>4.5</td>
<td>1.29</td>
</tr>
<tr>
<td>Delayed branch</td>
<td>0.5</td>
<td>1.07</td>
<td>4.6</td>
<td>1.31</td>
</tr>
</tbody>
</table>

Conditional & Unconditional = 14%, 65% change PC
Pipelining Summary

• Just overlap tasks, and easy if tasks are independent
• Speed Up ≤ Pipeline Depth; if ideal CPI is 1, then:

\[
\text{Speedup} = \frac{\text{Pipeline Depth}}{1 + \text{Pipeline stall CPI}} \times \frac{\text{Clock Cycle Unpipelined}}{\text{Clock Cycle Pipelined}}
\]

• Hazards limit performance on computers:
  – Structural: need more HW resources
  – Data (RAW,WAR,WAW): need forwarding, compiler scheduling
  – Control: delayed branch, prediction
Levels of the Memory Hierarchy

**Capacity**
- **CPU Registers**
  - 100s Bytes
  - <10s ns

- **Cache**
  - K Bytes
  - 10-100 ns
  - $0.01-.001/bit

- **Main Memory**
  - M Bytes
  - 100ns-1us
  - $0.01-.001

- **Disk**
  - G Bytes
  - ms
  - $10^{-3} - 10^{-4}

- **Tape**
  - infinite
  - sec-min
  - $10^{-6}$

**Access Time**
- **CPU Registers**
  - <10s ns

- **Cache**
  - 10-100 ns

- **Main Memory**
  - 100ns-1us

- **Disk**
  - ms

**Cost**
- **CPU Registers**
  - <10s ns

- **Cache**
  - $0.01-.001/bit

- **Main Memory**
  - $0.01-.001

- **Disk**
  - $10^{-3} - 10^{-4}$

**Upper Level**
- **Registers**
- **Cache**
- **Memory**
- **Disk**
- **Tape**

**Lower Level**
- **Instr. Operands**
- **Blocks**
- **Pages**
- **Files**

**Staging Xfer Unit**
- prog./compiler
  - 1-8 bytes

- cache cntl
  - 8-128 bytes

- OS
  - 512-4K bytes

- user/operator
  - Mbytes

**faster**

**Larger**
The Principle of Locality

• The Principle of Locality:
  – Program access a relatively small portion of the address space at any instant of time.

• Two Different Types of Locality:
  – Temporal Locality (Locality in Time): If an item is referenced, it will tend to be referenced again soon.
  – Spatial Locality (Locality in Space): If an item is referenced, items whose addresses are close by tend to be referenced soon.
Memory Hierarchy: Terminology

- **Hit**: data appears in some block in the upper level (example: Block X)
  - **Hit Rate**: the fraction of memory access found in the upper level
  - **Hit Time**: Time to access the upper level which consists of RAM access time + Time to determine hit/miss

- **Miss**: data needs to be retrieve from a block in the lower level (Block Y)
  - **Miss Rate** = 1 - (Hit Rate)
  - **Miss Penalty**: Time to replace a block in the upper level + Time to deliver the block the processor

- **Hit Time << Miss Penalty**
Cache Measures

- **Hit rate**: fraction found in that level
  - So high that usually talk about **Miss rate**
  - Miss rate fallacy: as MIPS to CPU performance, miss rate to average memory access time in memory

- **Average memory-access time**
  \[= \text{Hit time} + \text{Miss rate} \times \text{Miss penalty} \text{ (ns or clocks)}\]

- **Miss penalty**: time to replace a block from lower level, including time to replace in CPU
  - **access time**: time to lower level
    \[= f(\text{lower level latency})\]
  - **transfer time**: time to transfer block
    \[= f(\text{BW upper \\ & lower})\]
**Simplest Cache: Direct Mapped**

<table>
<thead>
<tr>
<th>Memory Address</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>

- **Location 0 can be occupied by data from:**
  - Memory location 0, 4, 8, ... etc.
  - In general: any memory location whose 2 LSBs of the address are 0s
  - Address<1:0> => cache index

- **Which one should we place in the cache?**
- **How can we tell which one is in the cache?**
1 KB Direct Mapped Cache, 32B blocks

- For a $2^N$ byte cache:
  - The uppermost $(32 - N)$ bits are always the Cache Tag
  - The lowest $M$ bits are the Byte Select (Block Size = $2^M$)

```
<table>
<thead>
<tr>
<th>Cache Tag</th>
<th>Example: 0x50</th>
<th>Cache Index</th>
<th>Byte Select</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0x50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Stated as part of the cache “state”

```
<table>
<thead>
<tr>
<th>Valid Bit</th>
<th>Cache Tag</th>
<th>Cache Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0x50</td>
<td></td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Byte 31</th>
<th>**</th>
<th>Byte 1</th>
<th>Byte 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Byte 63</td>
<td>**</td>
<td>Byte 33</td>
<td>Byte 32</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Byte 1023</td>
<td>**</td>
<td>Byte 992</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td>31</td>
</tr>
</tbody>
</table>
```
Two-way Set Associative Cache

• **N-way set associative**: N entries for each Cache Index
  – N direct mapped caches operate in parallel

• **Example: Two-way set associative cache**
  – Cache Index selects a “set” from the cache
  – The two tags in the set are compared in parallel
  – Data is selected based on the tag result
Disadvantage of Set Associative Cache

- **N-way Set Associative Cache v. Direct Mapped Cache:**
  - N comparators vs. 1
  - Extra MUX delay for the data
  - Data comes AFTER Hit/Miss

- **In a direct mapped cache, Cache Block is available BEFORE Hit/Miss:**
  - Possible to assume a hit and continue. Recover later if miss.
4 Questions for Memory Hierarchy

- Q1: Where can a block be placed in the upper level? (*Block placement*)
- Q2: How is a block found if it is in the upper level? (*Block identification*)
- Q3: Which block should be replaced on a miss? (*Block replacement*)
- Q4: What happens on a write? (*Write strategy*)
Q1: Where can a block be placed in the upper level?

- Block 12 placed in 8 block cache:
  - Fully associative, direct mapped, 2-way set associative
  - S.A. Mapping = Block Number Modulo Number Sets

![Diagram showing block placement in cache and memory]

DAP.F96 45
Q2: How is a block found if it is in the upper level?

- Tag on each block
  - No need to check index or block offset
- Increasing associativity shrinks index, expands tag
Q3: Which block should be replaced on a miss?

- Easy for Direct Mapped
- Set Associative or Fully Associative:
  - Random
  - LRU (Least Recently Used)

<table>
<thead>
<tr>
<th>Associativity:</th>
<th>2-way</th>
<th>4-way</th>
<th>8-way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>LRURandom</td>
<td>LRURandom</td>
<td>LRURandom</td>
</tr>
<tr>
<td>16 KB</td>
<td>5.2%</td>
<td>5.7%</td>
<td>4.7%</td>
</tr>
<tr>
<td>64 KB</td>
<td>1.9%</td>
<td>2.0%</td>
<td>1.5%</td>
</tr>
<tr>
<td>256 KB</td>
<td>1.15%</td>
<td>1.17%</td>
<td>1.13%</td>
</tr>
</tbody>
</table>
Q4: What happens on a write?

- **Write through**—The information is written to both the block in the cache and to the block in the lower-level memory.

- **Write back**—The information is written only to the block in the cache. The modified cache block is written to main memory only when it is replaced.
  - is block clean or dirty?

- Pros and Cons of each?
  - WT: read misses cannot result in writes
  - WB: no writes of repeated writes

- WT always combined with write buffers so that don’t wait for lower level memory
A Write Buffer is needed between the Cache and Memory

- Processor: writes data into the cache and the write buffer
- Memory controller: write contents of the buffer to memory

Write buffer is just a FIFO:

- Typical number of entries: 4
- Works fine if: Store frequency (w.r.t. time) \ll 1 / DRAM write cycle

Memory system designer’s nightmare:

- Store frequency (w.r.t. time) \rightarrow 1 / DRAM write cycle
- Write buffer saturation
5 minute Class Break

- 80 minutes straight is too long for me to lecture (12:40:00 – 2:00:00):
  - ≈ 1 minute: review last time & motivate this lecture
  - ≈ 20 minute lecture
  - ≈ 3 minutes: discuss class management
  - ≈ 25 minutes: lecture
  - 5 minutes: break
  - ≈ 25 minutes: lecture
  - ≈ 1 minute: summary of today’s important topics
A Modern Memory Hierarchy

- By taking advantage of the principle of locality:
  - Present the user with as much memory as is available in the cheapest technology.
  - Provide access at the speed offered by the fastest technology.

<table>
<thead>
<tr>
<th></th>
<th>Speed (ns)</th>
<th>Size (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>1s</td>
<td>100s</td>
</tr>
<tr>
<td>Control</td>
<td>10s</td>
<td>Ks</td>
</tr>
<tr>
<td>Datapath</td>
<td>100s</td>
<td>Ms</td>
</tr>
<tr>
<td>Registers</td>
<td>100s</td>
<td>Gs</td>
</tr>
<tr>
<td>On-Chip</td>
<td>100s</td>
<td>Tₛ</td>
</tr>
<tr>
<td>Cache</td>
<td>100s</td>
<td>10,000,000s</td>
</tr>
<tr>
<td>Second Level Cache (SRAM)</td>
<td>10s (ms)</td>
<td>10,000,000,000s (10s sec)</td>
</tr>
<tr>
<td>Main Memory (DRAM)</td>
<td>10,000,000s</td>
<td>10,000,000,000s</td>
</tr>
<tr>
<td>Secondary Storage (Disk)</td>
<td>10,000,000,000s</td>
<td>10,000,000,000s</td>
</tr>
<tr>
<td>Tertiary Storage (Disk)</td>
<td>10,000,000,000s</td>
<td>10,000,000,000s</td>
</tr>
</tbody>
</table>
Basic Issues in VM System Design

- size of information blocks that are transferred from secondary to main storage (M)
- block of information brought into M, and M is full, then some region of M must be released to make room for the new block --> replacement policy
- which region of M is to hold the new block --> placement policy
- missing item fetched from secondary memory only on the occurrence of a fault --> demand load policy

Paging Organization

virtual and physical address space partitioned into blocks of equal size

page frames
Address Map

\[ V = \{0, 1, \ldots, n - 1\} \] virtual address space
\[ M = \{0, 1, \ldots, m - 1\} \] physical address space
\[ n > m \]

MAP: \( V \rightarrow M \cup \emptyset \) address mapping function

\[ MAP(a) = a' \text{ if data at virtual address } a \text{ is present in physical address space} \]
\[ = \emptyset \text{ if data at virtual address } a \text{ is not present in } M \]
Paging Organization

Physical Memory

<table>
<thead>
<tr>
<th>Frame No.</th>
<th>Physical Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>7168</td>
<td>7168</td>
</tr>
</tbody>
</table>

Virtual Memory

<table>
<thead>
<tr>
<th>Page No.</th>
<th>Virtual Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>31744</td>
<td>31744</td>
</tr>
</tbody>
</table>

Address Mapping

Page Table

<table>
<thead>
<tr>
<th>Page Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page No.</td>
</tr>
<tr>
<td>Disp.</td>
</tr>
</tbody>
</table>

Page Table Base Register

Page Table index into page table

Virtual Memory Address (VA)

 Physically located in physical memory

Page Table

<table>
<thead>
<tr>
<th>Page Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
</tr>
<tr>
<td>Access Rights</td>
</tr>
<tr>
<td>PA</td>
</tr>
</tbody>
</table>

Virtual Memory Address (V.A.)

Physical Memory Address (P.A.)

Unit of mapping

Also unit of transfer from virtual to physical memory

Actually, concatenation is more likely
Virtual Address and a Cache

It takes an extra memory access to translate VA to PA

This makes cache access very expensive, and this is the "innermost loop" that you want to go as fast as possible

ASIDE: Why access cache with PA at all? VA caches have a problem!

*synonym / alias problem*: two different virtual addresses map to same physical address => two different cache entries holding data for the same physical address!

for update: must update all cache entries with same physical address or memory becomes inconsistent

determining this requires significant hardware, essentially an associative lookup on the physical address tags to see if you have multiple hits; or

software enforced *alias boundary*: same lsb of VA & PA > cache size
A way to speed up translation is to use a special cache of recently used page table entries -- this has many names, but the most frequently used is *Translation Lookaside Buffer* or *TLB*

<table>
<thead>
<tr>
<th>Virtual Address</th>
<th>Physical Address</th>
<th>Dirty</th>
<th>Ref</th>
<th>Valid</th>
<th>Access</th>
</tr>
</thead>
</table>

TLB access time comparable to cache access time (much less than main memory access time)
Translation Look-Aside Buffers

Just like any other cache, the TLB can be organized as fully associative, set associative, or direct mapped.

TLBs are usually small, typically not more than 128 - 256 entries even on high end machines. This permits fully associative lookup on these machines. Most mid-range machines use small n-way set associative organizations.
Reducing Translation Time

Machines with TLBs go one step further to reduce # cycles/cache access

They overlap the cache access with the TLB access

Works because high order bits of the VA are used to look in the TLB while low order bits are used as index into cache
Overlapped Cache & TLB Access

IF cache hit AND (cache tag = PA) then deliver data to CPU
ELSE IF cache miss and TLB hit THEN
   access memory with the PA from the TLB
ELSE do standard VA translation
Problems With Overlapped TLB Access

Overlapped access only works as long as the address bits used to index into the cache *do not change* as the result of VA translation.

This usually limits things to small caches, large page sizes, or high n-way set associative caches if you want a large cache.

Example: suppose everything the same except that the cache is increased to 8 K bytes instead of 4 K:

```
 11 2
cache index 00
20 12
virt page # disp
This bit is changed by VA translation, but is needed for cache lookup
```

Solutions:
go to 8K byte page sizes;
go to 2 way set associative cache; or
Summary #1:

• The Principle of Locality:
  – Program access a relatively small portion of the address space at any instant of time.
    » Temporal Locality: Locality in Time
    » Spatial Locality: Locality in Space

• Three Major Categories of Cache Misses:
  – Compulsory Misses: sad facts of life. Example: cold start misses.
  – Capacity Misses: increase cache size
  – Conflict Misses: increase cache size and/or associativity.
    Nightmare Scenario: ping pong effect!

• Write Policy:
  – Write Through: need a write buffer. Nightmare: WB saturation
  – Write Back: control can be complex
Summary #2: TLB, Virtual Memory

• Caches, TLBs, Virtual Memory all understood by examining how they deal with 4 questions: 1) Where can block be placed? 2) How is block found? 3) What block is replaced on miss? 4) How are writes handled?

• Page tables map virtual address to physical address

• TLBs are important for fast translation

• TLB misses are significant in processor performance: (funny times, as most systems can’t access all of 2nd level cache without TLB misses!)
Summary #3: Memory Hierachy

- Virtual memory was controversial at the time: can SW automatically manage 64KB across many programs?
  - 1000X DRAM growth removed the controversy
- Today VM allows many processes to share single memory without having to swap all processes to disk; VM protection is more important than memory hierarchy
- Today CPU time is a function of (ops, cache misses) vs. just f(ops): What does this mean to Compilers, Data structures, Algorithms?